A METHOD FOR AVOIDING F0 DISCONTINUITIES IN A CONCATENATIVE INTONATION MODEL

Francisco Campillo[*], *Taniya Mishra*[†], *Jan van Santen*[†]
Eduardo R. Banga[*]

[*] Signal Theory Group
Dpto. Teoría de la Señal y Comunicaciones
University of Vigo, SPAIN
[†] Center for Spoken Language Understanding
OGI School of Science & Engineering
20000 NW Walker Road, Beaverton, OR 97006, USA.
{campillo, erbanga}@gts.tsc.uvigo.es, {mishra, vansanten}@cslu.ogi.edu

ABSTRACT

This paper presents a novel approach for avoiding f0 discontinuities in the framework of a unit selection model of intonation with the accent group contour as the basic unit for concatenation. Taking as input the resulting contour of the model, it is modified on an accent group basis according to the typical decomposition into phrase and accent curves of the superpositional models. Results are presented that show a good performance of the proposed method, what implies that it can be a good approach for the treatment of the discontinuities, the main drawback of the concatenative models.

1. INTRODUCTION

Intonation modeling is widely acknowledged as one of the main factors regarding to the final quality of a TTS (Text-to-speech) system, independently of the underlying technology for generating the synthetic speech. This way, most current speech synthesizers produce a highly intelligible speech, but their naturalness is still far from being the desired one. In the literature we can find very different approaches to the problem of intonation, from the simple implicit modeling of [1], where the own intonation of the acoustic units resulting from unit selection is accepted, to other more complex methods. Particularly, in this paper we will focus on two classes: the superpositional and the concatenative models.

In the superpositional models ([2], [3]) the pitch curve is decomposed into several contributions corresponding to phonological entities with different temporal scope. The type of component curves, as well as the underlying assumptions on their possible shapes, make the distinction among the different superpositional models. For example, the Fujisaki model ([2]) considers three components: phrase curves, accent curves and a baseline constant value. However, the Linear Alignment Model ([3]) considers phrase curves, accent curves, and perturbation or residual curves. Independently of the set of curves, it is necessary to find an automatic method for decomposing the pitch contour to obtain every component, as well as a way for combining all this information and generating a target synthetic contour.

The concatenative models ([4], [5], [6]) do not make any assumption on the nature of the intonation, and try to generate the pitch contour by concatenating natural contours extracted from similar contexts to those where they are applied, in a process identical to the acoustic unit selection. These models are very simple and the task of analyzing the natural pitch contour is reduced to chunking it into consecutive sections according to the temporal scope of the phonological entity the basic unit is tied to. However, this models have an important flaw. If the available intonation corpus is not large enough, or the unit selection algorithm is not properly designed, there can be discontinuities in the concatenation points, what severely degrades the performance of the model. In this paper we will address this problem, proposing a novel method that employs the different components assumed by the superpositional models for eliminating the remaining discontinuities.

The outline of this paper is as follows. Sections 2 and 3 describe the main characteristics of the superpositional and concatenative models, respectively. Then, section 4 presents a comparison of both models, emphasizing their common points and their differences. Section 5 shows a method for modifying the pitch contour of an accent group in order to avoid the possible f0 discontinuities in the points of concatenation with the surrounding accent groups. Sections 6 and 7 describe a perceptual test conducted to check the performance of the proposed method, and a discussion of the results. Finally, section 8 is dedicated to the conclusions and some suggestions for future research.

2. SUPERPOSITIONAL MODELS

The superpositional models consider the pitch curve as the contribution of several component curves related to different phonological domains. For example, in the Fujisaki model ([2]), the most known superpositional model, the pitch curve is the result of the addition in the logarithmic domain of phrase curves (tied to the phonological phrase), accent curves (tied to the accented syllables) and a constant minimum level dependent on the speaker. Another example, the General Superpositional Model of Intonation ([3]), assumes the pitch curve to be the result of the addition in the log domain of three different components: the phrase and accent curves, and a new one, the residual or perturbation curve, related to the phonemes, which permits a more natural approximation of the pitch curve on taking into account the own microprosody of the segments. With respect to the previous model, where very hard constraints are held into the possible shape of the curves, in the General Superpositional Model these shapes are not specified. As a result, the analysis and extraction of the different components is a very hard task,

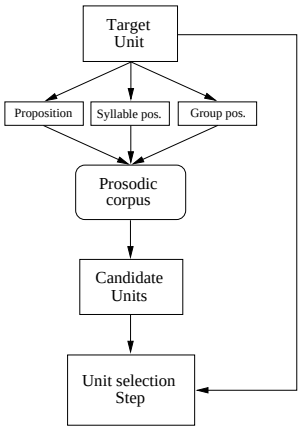

Figura 1. *Clustering of accent groups in the intonation unit selection*

although there are several promising approaches ([7], [8]).

3. CONCATENATIVE MODELS

Based on the same principles as the acoustic unit concatenation systems, these methods try to model the pitch contour as the result of the concatenation of natural contours. The underlying assumption is that an intonation unit applied in a very similar context to its original one should be almost indistinguishable from a natural contour. There are several approaches ([5], [6], [4]) with differences in the basic units for concatenation and the features modeling the context, but in this paper we will focus solely on the first one.

In [5] the process for generating the global pitch contour is formally equivalent to the acoustic unit selection of [9], which means that it is necessary to decide a basic unit for concatenation, a way of choosing a subset of candidate units for every target unit, and design a target and a concatenation cost functions for unit selection. In this case the basic intonational unit is the accent group, defined as a sequence of non accented words ending in an accented word. The subset of candidate groups is decided according to previous knowledge of the Spanish intonation ([10]): the accent groups are clustered into 48 classes as a function of three features: the position of the accent group within the phonic group (initial, final, medial and initial-final), the type of proposition (declarative, interrogative, exclamatory and ellipsis) and the position of the accented syllable within the accent group (ultimate, penultimate and antepenultimate). Figure 1 shows a diagram of this first hard constraint that decides the subset of candidate units for a unit selection step. After that, the target cost function considers other features, such as the number of syllables, the duration of the accent group, the position of the phonic group within the sentence and the type of break at the end of the group, while the concatenation cost measures mainly the continuity of f0 at the joint point. Finally, the sequence of accent groups with the best score taking into account all the target and concatenation costs is selected.

4. COMPARISON OF METHODS

Comparing two different approaches describing the same phenomenon is always interesting, as it can provide some very useful insight on the nature of the problem. In the superpositional models it is necessary to decompose the global pitch contour into the different contributions at analysis time, which, as mentioned before, it is not a trivial problem. At synthesis time, how to find the best combination of phrase curves and accent curves is still an open question, although there are interesting approaches like [11].

In a concatenative system the analysis of the pitch contour is a trivial problem, as you only need to find the join points, what depends directly on the chosen unit. At synthesis time it is necessary to find a suitable set of features for describing the context (what it is likely to be crucial for every intonation model), and the design of the cost functions for selection, which introduces the currently unsolved problem of the weight optimization of every subcost contribution to the global score.

The accent group contour chosen in [5] can be considered as the addition of an accent curve and a local phrase curve, in terms of the superpositional models. The clustering shown in Figure 1, that assures initial and final accent groups for the beginning and end of the phrase, respectively, implies an implicit phrase curve selection, with the intermediate groups looking for continuity with the concatenation cost function. In fact, in [3] the phrase curve is modeled as a two-part curve obtained by non linear interpolation between three points: the start of the phrase, the start of the last accent group in the phrase, and the end of the phrase. So, on considering these points in the clustering, an appropriate declination according to the type of sentence is chosen, avoiding the need for decomposing the original contour and finally adding the different contributions in a process very likely to be prone to errors.

However, there is an important flaw in the concatenative methods: their behavior in the presence of discontinuities. In a superpositional framework this problem does not exit, as the explicit use of the phrase curve assures continuity in every point of the pitch contour. The intonation unit selection approach in [5] minimizes this problem on considering multiple candidates for every target unit, and it has shown a very good performance in the domain of neutral speech with a reasonable sized prosodic corpus (about one hour), but the problem of infrequent pitch jumps is potentially still there. Moreover, it can be argued that it is just a problem of corpus coverage, but the application to other domains such as expressive speech, where the larger variability would likely imply more discontinuities, makes it important to research some method for treating these cases where the unit selection algorithm can not find a suitable sequence of intonation units.

5. PITCH CONTOUR MODIFICATION

Summing up, it would be desirable to find a method that uses natural units as often as possible, while avoiding discontinuities. As mentioned before, the accent group employed in [5] includes both phrase and accent curves from the superpositional model. This suggests the possibility of combining the basics of both methods, performing a unit selection for obtaining the better sequence of natural contours and applying the general assumption of the superpositional models of the different components for fixing the existing discontinuities. So, in the presence of a discontinuity, we could maintain the accent curve component of the chosen accent group, and add it to a new local phrase curve computed in order to balance the discontinuity.

This approach would solve the problem of the discontinuities, but we would still need to decompose the candidate pitch contours. Moreover, now we would also have to make a local phrase curve estimation from the implicit one defined by the surrounding accent groups.

In [8] it is assumed that the phrase curve can be approxi-

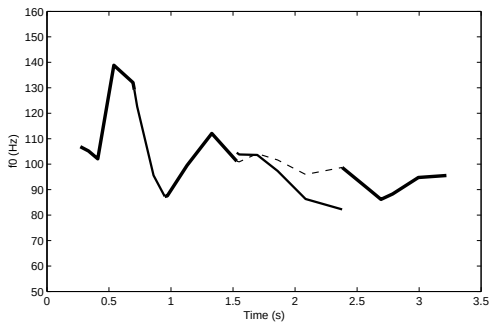

Figura 2. *Modification of an accent group pitch contour*

mated by n line segments, where n is the number of feet in the phrase. Although the left-headed foot used there is defined as an accented syllable followed by one or more unaccented syllables, and therefore its temporal scope does not coincide with the accent group defined in section 3, we will still assume this approximation to be valid for our case. Then, if the phrase curve is a line segment in the scope of every accent group, the pitch contour modification can be extremely easy. First, if the intonation unit selection algorithm is well designed, and the prosodic corpus can provide a sequence of accent groups without very large discontinuities between groups, the desired local phrase curve can be reasonably approximated by the segment line joining the desired two points of concatenation with the surrounding contours. And second, being the original phrase curve of the accent group a segment line, and given that the addition of two segment lines is another segment line, the pitch contour modification is reduced to subtract from the accent group the segment line that compensates the differences in f0 in both concatenation points. In other words, it is not necessary to compute the target local phrase curve and the original phrase curve.

Equations (1), (2) and (3) show the algorithm for computing the pitch contour modification $\Delta(t)$ as a function of the differences $D_{initial}$ and D_{final} at times $t_{initial}$ and t_{final}, corresponding to the discontinuities at the beginning and end of the accent group. The differences $D_{initial}$ and D_{final} are computed subtracting the original value of the accent group from the desired value at the concatenation points.

$$\Delta(t) = m \times t + b \tag{1}$$

where the slope m is

$$m = \frac{D_{final} - D_{initial}}{t_{final} - t_{initial}} \tag{2}$$

and the intercept b

$$b = D_{initial} - m \times t_{initial} \tag{3}$$

Then, the pitch contour f_0 of the accent group is modified in the following manner:

$$f(t) = f_0(t) + \Delta(t) \tag{4}$$

Figure 2 shows an example of the application of this modification method. The pitch contour resulting from the intonation unit selection is represented by a solid line. The accent group located around 1.5 and 2.5 seconds exhibits a discontinuity of about -4 Hz at the beginning and 15 Hz at the end. The modified contour is represented by a dotted line.

6. PERCEPTUAL TESTS

A listening test was conducted in order to check the performance of the modification procedure. 30 sentences were manually extracted from our Galician speech corpus, looking for accent groups belonging to the same class (see Figure 1) and with similar context (number of syllables, position of the accent group into the phonic group, number of words, etc), and big discontinuities $D_{initial}$ and D_{final} in the points of concatenation. From this set, 12 sentences were declarative, 10 interrogative, 6 exclamatory and 2 ellipsis. Table 1 shows the statistics of the discontinuities $D_{initial}$ and D_{final} before modification, while the $Slope$ term is the difference between both discontinuities divided by the duration of the new contour in its original context, and gives an idea of the rate of change.

Tabla 1. *Perceptual test sentences statistics*

	Mean	Std. Dev	$\|Max\|$	$\|Min\|$
$D_{initial}\ (Hz)$	-4.53	14.18	30.97	2.27
$D_{final}\ (Hz)$	-2.78	12.45	21.39	4.72
$Slope\ (Hz/s)$	-2.04	32.55	114.37	2.62

For every sentence two stimuli were obtained as the output of a unit-selection speech synthesizer ([12], [13]), the first one imposing its original intonation contour, and the second one imposing the new one resulting from the change of an accent group contour and the subsequent application of the proposed modification method.

The perceptual test was a typical CMOS (Comparison Mean Opinion Score) where every listener was asked to evaluate each pair of stimuli according to a 5 point scale, ranging from "Stimulus A much better than B" to "Stimulus B much better than A", taking into account only the naturalness of the synthetic intonation curves (Table 2). The order of presentation of every pair of stimuli was randomized, as well as the stimuli themselves within each pair. 14 native or bilingual listeners participated in the test, 6 of them involved in speech research.

1	Stimulus A much better than B
2	Stimulus A better than B
3	Similar or equal
4	Stimulus B better than A
5	Stimulus B much better than A

Tabla 2. Perceptual test scale

7. DISCUSSION

Figure 3 shows the results of the perceptual test, taking the natural contours as the reference. About 80% of the evaluations considered that the modified stimuli were not at least worst than the original ones, which is a very good result if we take into account that the modification method had to deal with discontinuities as large as 30 Hz in some cases (see Table 1). Although the differences between the percentages of evaluations considering worse (including much worse) and better (including much better) were found to be statistically significant after performing a one-sample binomial proportion test (p-value = 0.0133), the authors argue that some of the differences found in the test can be explained by listeners' evaluation based more on their own preferences than on the naturalness of the synthetic contours, as suggest the funny result of about 11% of modified contours being more natural than the original ones. Moreover, another source of

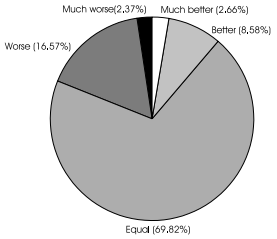

Figura 3. *Results from the perceptual test*

distortion comes from the use of a corpus based speech system for obtaining the stimuli. On using the original contours for the acoustic unit selection it was very likely to get the sequence of units extracted from that same sentence, producing a synthetic stimulus with a higher global quality than the stimulus with the modified contour. Finally, with respect to the sentences, only the modified versions of 2 of the remaining 30 were considered to be worse by the majority of the listeners.

8. CONCLUSIONS

In this paper we presented a novel method for modifying the pitch contour of an accent group, in order to avoid the discontinuities that degrade drastically the naturalness of the global synthetic curve. With this goal in mind, we combined the output of an intonation unit selection system ([5]) with the typical decomposition of the superpositional models ([3]), and proposed a local modification of the accent group phrase curve for eliminating the discontinuities. This way, natural contours are used as far as possible, and some of them are modified when required without having to decompose them explicitly into their different components.

The possibility of modifying the pitch contours gives more freedom to a unit selection intonation model, because the f0 continuity is a very hard constraint in the design and optimization of the cost functions. Nevertheless, it is still an important factor, as the local phrase curve estimation described in section 5 requires some continuity in the sequence of chosen accent group contours to be a reasonable approximation. Anyway, this point deserves more research effort.

Note that the choice of the accent group as the basic unit permits this simple modification technique. For instance, smoothing discontinuities using the segment instead ([6]) would possibly be much harder.

The subjective results showed a very good performance of the modification method, even in the presence of quite large discontinuities. Although different types of sentences were included in the test, looking for variability, the experiment was restricted to neutral speech. A similar perceptual test with expressive speech remains as a very interesting future work. A good behavior of the modification method there would present the unit selection technique as a very good candidate for intonation modeling for every domain and without relying on the availability of a very large prosodic corpus.

9. ACKNOWLEDGMENTS

The work reported here was carried out while a visiting research post doc at Center for Spoken Language Understanding of the first author, with funds from Dirección Xeral de Investigación, Desenvolvemento e Innovación, Consellería de Innovación e Industria, Xunta de Galicia, and support from NSF grant 0205731 "ITR: Prosody Generation for Child Oriented Speech Synthesis" to Jan van Santen.

10. REFERENCES

[1] Black, A. and Taylor, P. "Automatically clustering similar units for unit selection in speech synthesis". In Proceedings of EUROSPEECH'97, p601–604, Rhodes, 1997.

[2] Fujisaki, H., "Dynamic characteristics of voice fundamental frequency in speech and singing". In Peter MacNeilage, editor, The Production of Speech, p39–55, Springer, New York, NY, 1983.

[3] van Santen, J., and Möbius, B., "A quantitative model of f0 generation and alignment". In Botinis, editor, Intonation Analysis, Modelling and Technology, chapter 12, p269–288, Kluwer Academic Publishers, Netherlands, 1999.

[4] Meron, J., "Prosodic unit selection using an imitation speech database". In 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland, 2001.

[5] Campillo, F. and R. Banga, E., "Combined prosody and unit selection for Corpus-based text-to-speech systems". In Proceedings of ICSLP'02, p141–144, Denver, 2002.

[6] Raux, A. and Black, A., "A unit selection approach to F0 modeling and its applications to emphasis". In Proceedings of ASRU 2003, St Thomas, US Virgin Islands, 2003.

[7] van Santen, J., Mishra, T. and Klabbers, E., "Estimating phrase curves in the general superpositional intonation model". In Proceedings of the ISCA Speech Synthesis Workshop'04, Pittsburgh, PA, 2004.

[8] Mishra, T., van Santen, J. and Klabbers, E., "Decomposition of pitch curves in the general superpositional intonation model". Prosody 2006, Dresden, Germany, 2006.

[9] Black, A. and Campbell, N., "Optimising selection of units from speech databases for concatenative synthesis". In Proceedings of EUROSPEECH'95, p581–584, Madrid, 1995.

[10] López, E., "Estudio de técnicas de procesado lingüístico y acústico para sistemas de conversión texto voz en Español basados en concatenación de unidades". PhD thesis, E.T.S.I. de Telecomunicaciones, Universidad Politécnica de Madrid, España, 1993.

[11] van Santen, J., Kain, A., Klabbers, E. and Mishra, T. "Synthesis of prosody using multi-level unit sequences". In Speech Communication, Volume 46, p365–375, 2005.

[12] Campillo, F. and Banga, E. R., "On the design of the cost functions for a unit selection speech synthesis". In Proceedings of EUROSPEECH'03, p289–292, Geneva, 2003.

[13] Campillo, F. and R. Banga, E., "A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems". Speech Communication 48, pag. 941-956, 2006.